An Exploratory Study of Social Media Analysis for Rare Diseases using Machine Learning Algorithms: A case study of Trigeminal Neuralgia

Haadi Mombini
Worcester Polytechnic Institute
hmombini@wpi.edu

Ruojun Li
Worcester Polytechnic Institute
rli2@wpi.edu

Yixin Zhang
Worcester Polytechnic Institute
yzhang17@wpi.edu

Dmitry Korkin
Worcester Polytechnic Institute
dkorkin@wpi.edu

Bengisu Tulu
Worcester Polytechnic Institute
bengisu@wpi.edu

Abstract

Rare diseases, affecting approximately 30 million Americans, are often poorly understood by clinicians due to lack of familiarity with the disease and proper research. Patients with rare diseases are often unfavorably treated, especially those with extremely painful chronic orofacial rare disorders. In the absence of structured knowledge, such patients often choose social media to seek help from peers within patient-oriented social media communities thereby generating tremendous amounts of unstructured data daily. We investigate whether we can organize this unstructured data using machine learning to help members of rare communities find relevant information more efficiently in real-time. We chose Trigeminal Neuralgia (TN), an extremely painful rare disorder, as our case study and collected 20,000 social media TN posts. We categorized TN posts into Twitter (very short), and Facebook (short, medium, long) datasets based on message length and performed three clustering experiments. Results revealed GSDMM outperformed both K-means and Spherical K-means in clustering Facebook especially for short messages in terms of speed. For long messages, MDS reduction outperformed the PCA when both were used with K-means and Spherical K-means. Our study demonstrated the need for further topic modeling to utilize among high level clusters based on semantic analysis of posts within each cluster.

1. Introduction

Delivering the best quality of care equally to all patients is clinically ideal but practically challenging. It is, in many cases, due to lack of specialty knowledge and expertise [1, 2]. This challenge is more severe for patients, clinicians and other stakeholders facing rare diseases [1]. A rare disease is defined as an incident that affects fewer than 200,000 people in the United States at any given time [3]. Rare diseases affect an estimated 25 million to 30 million Americans [3]. They are often difficult to diagnose [1] and poorly understood by clinicians due to lack of familiarity with, or even basic awareness of the disease [1, 4]. As a result, little research is being performed in many of them, leading to slow advances in clinical care, limited confidence in both diagnosis and management [1, 2] and low quality of life.

One of the extremely painful [5] rare chronic orofacial diseases is called Trigeminal Neuralgia (TN) which is diagnosed in 150,000 people each year [6]. TN is rare in pain experience [7], clinical signs and symptoms [5], making its management extremely challenging for patients, clinicians and stakeholders. In view of the rarity of TN, few general practitioners have experience dealing with TN patients [2]. For these practitioners, there are several different classifications and definitions to guide them (classic, idiopathic, secondary, and symptomatic TN) thereby leading to diagnosis confusion [8]. Considering complex management [9], medications with serious side effects [10] and symptoms that are frequently mistaken for dental or jaw pain [8, 11], TN patients may face psychological issues leading to mild to severe functional limitation of daily life activities [12], which - based on studies in Europe [2]- could lead to suicide. The incident is more prevalent among women older than 40 years old [13]; however, studies reported cases before age 20 in 1% -1.5% of patients [14].

Due to the lack of efficient clinical and diagnostic procedures and difficulty of obtaining required clinical data, patients with rare disease (including TN) and clinicians have gathered in online social health communities (usually Facebook and Twitter) to interact and discuss such diseases [15-18]. This information exchange between the social media users however, according to recent reviews [19, 20], needs to be monitored for quality and reliability.

URI: https://hdl.handle.net/10125/64211
978-0-9981331-3-3

Typically, within online social media health and disease groups (e.g. lung cancer) the quality of information may be assured by either the experienced party sharing credible source or the inexperienced party studying the topic further. Due to limited experience and lack of structured knowledge in TN social media communities, this practice of quality assurance of information is limited. Although, there is an increasing rate for physician participation within social networking and microblog sites (i.e. Facebook and Twitter) [19, 21], TN social media posts (as shown in Figure 1) show that majority of information exchange occurs between novice individuals. This puts TN social media data exchange at high risk of low information quality. Hence, TN differs from other social health communities in terms of social and functional building blocks, as suggested in social ecology framework [20, 22]. In particular, TN may differ in the extent to which patients (1) reveal their identity, (2) communicate with each other, (3) share content, (4) know if others are present to help, (5) relate to each other, (6) know of TN content, and (7) form such communities [22].

Among these social blocks, what makes TN social media outstanding is the high level of content sharing and openness as well as the level of knowledge about TN content among patients which have been similarly reported for some other rare diseases [18]. Given that this information exchange among TN patients generate tremendous amounts of unstructured data (Figure 1), we investigate whether this unstructured data can be efficiently organized so that the TN community can benefit from more structured information presentation within online communities. In this study, we identify ways to explore and organize the unstructured data.

Using our dataset of 20,000 TN posts collected from Facebook and Twitter, we 1) apply text preparation methods to utilize the social media posts from the dataset, 2) apply clustering algorithms and run different experiments to identify the most suitable patterns and structures for TN online community posts and compare and contrast such algorithms that are more efficient for these types of posts and finally 3) give insights for application of supervised machine learning algorithms to take advantage of those clusters that are more meaningful so that they can be classified into different categories.

The result of this study can open new research opportunities towards better utilization of patient generated health information among online communities for rare diseases such as TN.

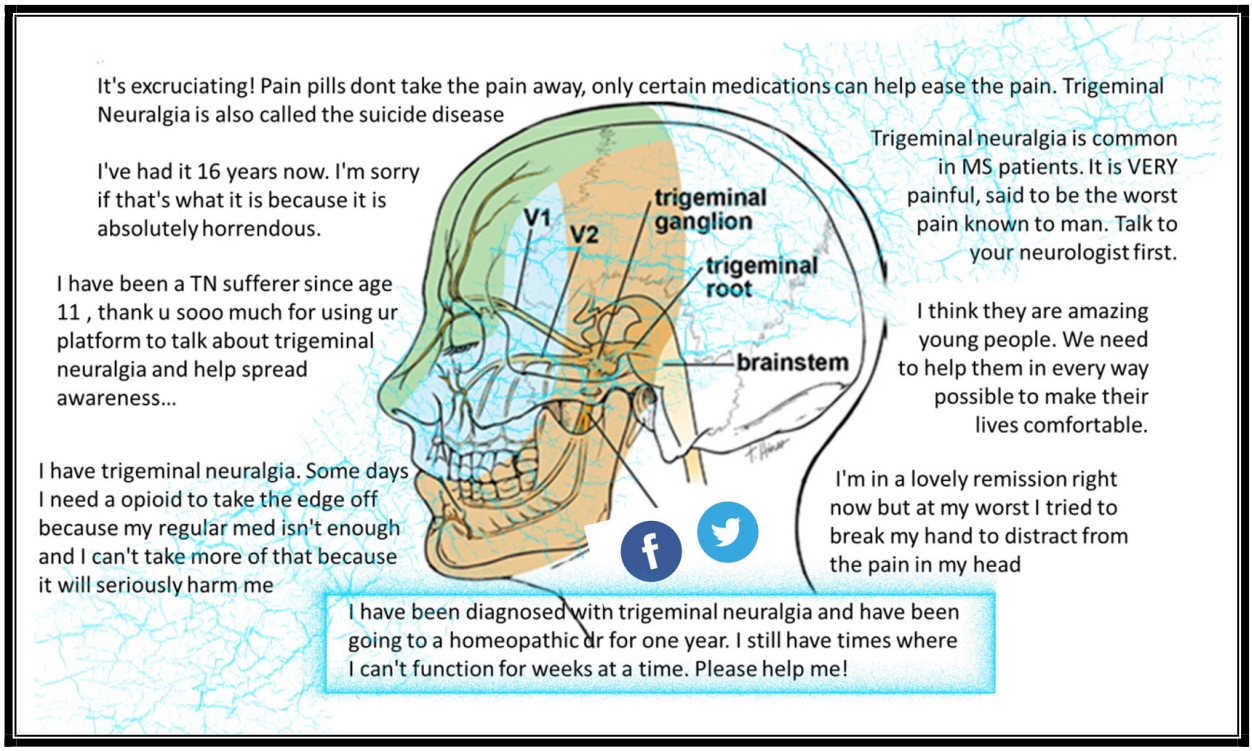

Figure 1. Sample unstructured posts by TN patients from Facebook dataset (Image is from fpa-support.org)

2. Background

Social media platforms such as Facebook and Twitter became popular sources of knowledge discovery about specific diseases and patients' online social engagement [23-28]. In case of rare diseases, these platforms may be the only source of knowledge for both patients [29] and physicians [30, 31] because in those areas often large sets of (real patient medical) data for research and analysis (e.g. training machine learning algorithms) are missing [32]. Therefore, physicians may seek help from a social health community to decide on the most suitable treatment plans [31]. Genetics companies are also able to hire patients more effectively through social media sites for critical research on rare diseases [31]. Given the wealth of knowledge accumulated in social media communities of rare diseases, analyzing the content using machine learning techniques can provide new knowledge and research directions at a low cost. However, development of analytical approaches for rare diseases are challenging [33].

Studies that proposed machine learning analytical techniques for mining Facebook and Twitter health related data to help with the management of rare diseases are limited. Although proposed social media analytics demonstrated initial success, its use for improving health related research is still at its early stages [31]. To the best of our knowledge, no study has yet focused on analyzing social media posts generated by TN patients.

Current methods practiced by studies are considered to be keyword-based and supervised-learning-based methods that are used to identify disease-related textual information from social media data [34]. In text classification, the idea is to find the best matching category for the text document [23]. To classify texts, several algorithms have been proposed. These algorithms are mostly used for clustering documents in an unsupervised fashion when data labels are not present which is the case with any social media data analysis [35]. In clustering problems, the idea is to find groups of similar objects in the data. The similarity between the objects is measured with the use of a similarity function [35]. Among many text weighting schemes explored, the term frequency-inverse document frequency (TD-IDF) is commonly used to weight each word in the text document according to how unique it is [23]. TF-IDF works by determining the relative frequency of words in a specific document compared to the inverse proportion of that word over the entire document corpus [36]. Among the classical clustering algorithms, Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model (GSDMM), K-means and the Spherical algorithms are widely used. GSDMM is useful for dealing with tweets and used by several studies [37-40]. Different variations of K-means and Spherical K-means were also used for rare disease analysis [41].

3. Methodology

3.1 Description of dataset

We collected a total of 9,808 tweets from Twitter and 10,000 public posts from Facebook using the help of Crimson Hexagon, a leading social media analysis software company. Both Twitter and Facebook data range from January 1st, 2015 to January 30th, 2018. The keyword combination for the search was simply chosen to be “TN” and “Trigeminal Neuralgia” to collect larger sets of posts.

3.2 Data preparation

As a first step, data preprocessing is needed before performing any document clustering. To make the most out of the dataset (approximately 20,000 data samples from both Twitter and Facebook), duplicate records and irrelevant posts such as advertisements, broken links, websites, and general news were identified and removed. Posts from each user were identified and then combined into one full set record for each user so that each user can be associated with a unique set of posts. This data preparation helps to ensure both variability and consistency of each user's contents in terms of the time of the post and different experience.

Next, we create a space of features that comprises a reference from which document vectors are selected [42]. Each feature references a term that occurs in the document collection [42]. We adopted data preparation scheme as follows:

Word Tokenization: Since our dataset contains short phrases (posts and tweets) and TN is rare and does not have a popular phrase library, we took advantage of word tokenization in which we divided phrases into tokens (words). In the experiments described later, we used Python NLTK library and specifically chose “word_tokenize()” function to perform tokenization task.

Stop words deletion: In computing, stop words are those that are filtered out before or after processing of natural language textual data or text [19]. For this analysis, we used stop words from “sklearn.NLTK.en” stop words library (which contains a set of words such as am, is, the, *etc*). Besides, we added extra stop words including "http," "rt' which are commonly used in social media.

Corpus normalization, Vocab creation and vectorization: In this step, we normalized and restricted the textual data to some limited words or tokens that are more useful among several others based on their distributions [43]. In general, this can be done using weighting vocabularies methods such as TF-IDF. For our experiments (described later), we used python "TfidfVectorizer" function. The dataset distribution based on text length after the general cleansing step is illustrated in Figure 2.

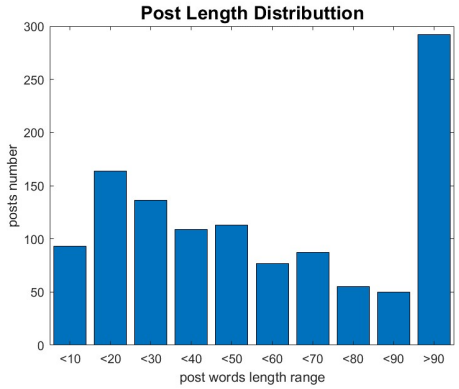

Figure 2. Frequency distribution and length of cleaned posts from patients on Facebook and Twitter combined.

To run our machine learning experiments using Facebook dataset, we set 50-word and 100-word thresholds and consider messages less than 50 words as short, between 50 to 100 words as medium and larger than 100 words as long messages. We categorized Twitter dataset as "very short" messages. Consequently, (after data cleansings and preparation steps) most of the distributions proved to be in the short message group and only about 250 posts contain more than 100 words which we consider as long messages. Therefore, the distribution is biased towards short messages (as shown in Figure 2 above).

Since we are dealing with unsupervised clustering tasks with unlabeled documents (posts and tweets) and TN does not have a prior standard set of vocabulary, we created a baseline vocabulary for our machine learning experiments using Bag-Of-Words model so we can compare the results of clustering algorithms with this baseline. In Bag-Of-Words, words are represented as sets of words, and the frequency of each word corresponds to a feature in the resulting multi-dimensional vector space. Thus, each document is then represented as a feature vector of non-negative values [44]. Words that appear more frequently will be valued as more critical and descriptive for the document. Since Bag-Of-Words model has limitations [44] for the short-text documents, we only use it to create baseline TN categories that we illustrate through WordCloud representation of common themes.

3.3 Machine learning approach

In this research, we choose from unsupervised clustering algorithms to make sense of our social media dataset of TN posts (Twitter, Facebook short, medium and long). We first begin by using bag of words which is widely used in text mining [34]. In this technique, words are assumed to appear independently, and the order is immaterial [34]. Then, a distance-based method is used based on term frequency and Inverse Document Frequency (TF-IDF) to analyze TN posts collected from Facebook and Twitter to help understand the structure of patients' contents. In distance-based clustering tasks based on similarity, choosing an appropriate similarity measure is critical for cluster analysis, especially for a particular type of clustering algorithms [45].

Cosine similarity based on term frequency (TF) is selected as the method's similarity measure. We choose to use this similarity measure due to its wide applicability specifically to clustering text documents [45]. We then run and compare the performance of several clustering algorithms to achieve the most efficient clustering results. These include Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model (GSDMM), K-Means and the Online-Spherical algorithms.

GSDMM is a popular clustering technique used for short text topic modelling [46] which is useful to deal with our short message posts and tweets. This algorithm has a good balance between the completeness and homogeneity of the clustering results and is fast to converge [47].

In K-means clustering data space is partitioned into k different clusters of objects, so that the sum of squared Euclidean distances between the center of each cluster and the individual objects inside that cluster is minimized [48]. The goodness-of-fit of K-means algorithm is often expressed in terms of amount of variance explained. The following formula presents this variability:

$$\mathbf{VAR_{TOT}} = \frac{\mathbf{SS_{TOT}} - \sum_{i=1}^{K} SS_i}{\mathbf{SS_{TOT}}} \qquad [48]$$

where SS_{TOT} is the total sum of squares in the data space and SS_i is the within sum of square of the i^{th} cluster. VAR_{TOT} is analogous to the conventional R^2 [48].

If K-means clustering uses the cosine similarity it is known as the spherical K-means algorithm [42]. It can be applied to document vectors or any type of directional data [42].

Making sense of social media data necessitates methods that can best represent the data. Since such data is sparse and often contain lots of textual features reducing the feature dimensions will help understand the data better. Therefore, dimension reduction techniques such as principal component analysis (PCA) and multidimensional scaling (MDS) will be used to make more sense of clusters.

3.4 Proposed experiments

3.4.1 Creating baseline vocabulary set. Although text clustering is a useful and inexpensive way to organize vast text repositories into meaningful topical categories, there is little consensus on which clustering techniques work best, and in what circumstances since researchers usually do not use the same evaluation methodologies and document collections [42]. The typical evaluation method is Normalized Mutual Information (NMI), Homogeneity (H), Completeness (C), Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), and topic coherence [49]. However, for this study, these methods need a test dataset (labeled by human raters [49]) to represent the disease severity which is not possible due to time and effort it requires. Therefore, to evaluate the clustering algorithms, we compared actual instances of Facebook posts and tweets clustered by each algorithm to that of baseline extracted using our Bag-Of-Words model (Figure 3).

Figure 3. WordCloud representation of Bag-Of-Words for top 100 words (Facebook and Twitter)

Using WordCloud representation of our Bag-Of-Words model, we were able to infer three baseline vocabulary sets as shown in Table 1. These baseline vocabularies were compared with those clustered by the algorithms so that we can make more sense of their performance in terms of thematic representation of TN related posts from both Facebook and Twitter.

Table 1. Baseline vocabulary sets

Baseline set	Representative words
Awareness (For world to know)	Awareness, day, hope, world, people, support, cure, understand, relief, please
Pain experience (For other sufferers to learn)	pain, meds, doctor, carbamazepine, back, used, suffering, try
TN Description (To define and differentiate from similar facial disorders)	Suicide, disease, cause, Fibromyalgia, many, treatment

Then, we ran three different experiments using GSDMM, K-means and Spherical K-means on four datasets (Twitter, Facebook short, Facebook medium and Facebook long) to determine how each of these classical clustering algorithms perform on different datasets. For a fair comparison between clustering algorithms, considering that the datasets are not equally distributed (only 255 samples for long messages exist), we selected a sample of 250 from both short and long message datasets.

To deal with high dimensionality of textual features within our dataset, the classical yet powerful techniques for dimensionality reduction, PCA and MDS were used. Both PCA and MDS are simple to implement, efficiently computable, and guaranteed to discover the true structure of data lying on or near a linear subspace of the high-dimensional input space [50]. PCA finds a low-dimensional embedding of the data points that best preserves their variance as measured in the high-dimensional input space whereas MDS finds an embedding that preserves the inter-point distances [50]. When using Euclidean distances MDS becomes equivalent to PCA [50]. After data cleansing step and applying normalization using TF-IDF, the final vocabulary matrix comprised of 14 features as illustrated in Figure 4.

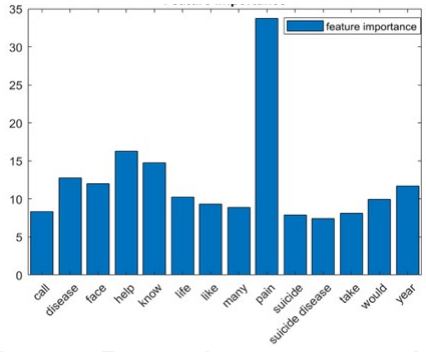

Figure 4. Feature importance matrix representing the most important words discussed among TN patients.

4. Results

We first applied GSDMM algorithm to cluster posts from all datasets (Figure 5). This algorithm is popular for clustering short text messages, which in our case are mostly tweets from our Twitter dataset.

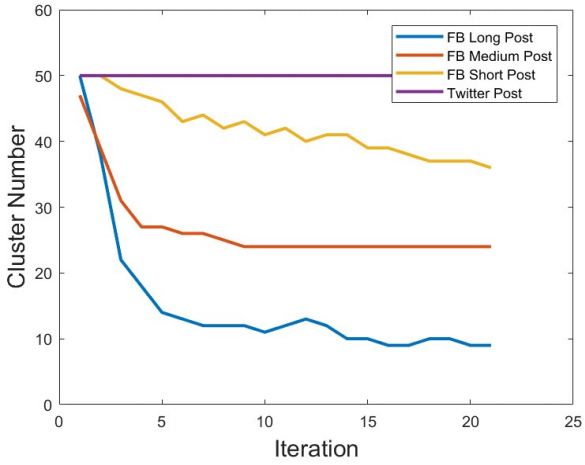

Figure 5. GSDMM cluster number and iteration number.

The GSDMM results on the Facebook and Twitter datasets show that the number of clusters is reduced after each iteration (Figure 5). GSDMM could cluster posts longer than 50 words with a good trend. When a post is too short (less than 50 words), even the GSDMM performs poorly and it takes 100 iterations. This signifies the nature and relationship between number of posts and clusters as presented in Table 2 below.

Table 2. GSDMM results for different post length (Facebook and Twitter)

Post Length (words)	50	100	150	200	250
Cluster Number	3	3	3	3	3
Iteration Number	100	80	26	20	2

According to Table 2, GSDMM could fit the Facebook dataset faster than Twitter dataset which is unexpected. One explanation could be that our Twitter data has less heterogeneity in terms of topics because TN patients mostly tweet to raise awareness for TN.

In our second and third experiments, we used dimension reduction techniques (MDS and PCA) and ran K-means and Spherical K-means algorithms respectively to cluster the TN messages from both Facebook and Twitter datasets (Figures 6 and 7).

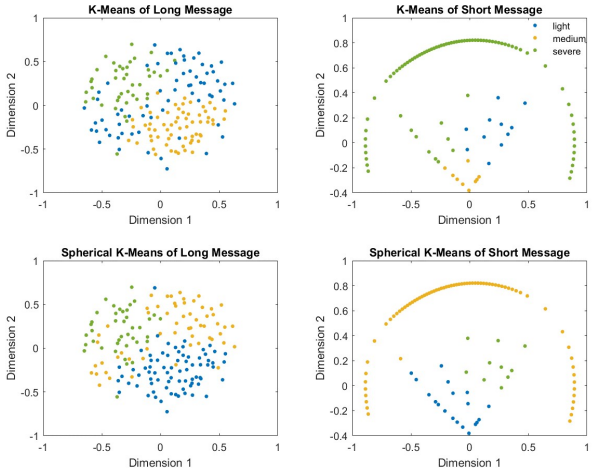

Figure 6. Clustering representation of K-means and Spherical K-means using cosine similarity with MDS.

Results of such experiments demonstrated that the classical K-means and Spherical K-means algorithms performed better in terms of clustering long message posts (Facebook dataset) as represented with the dimension-reductions techniques.

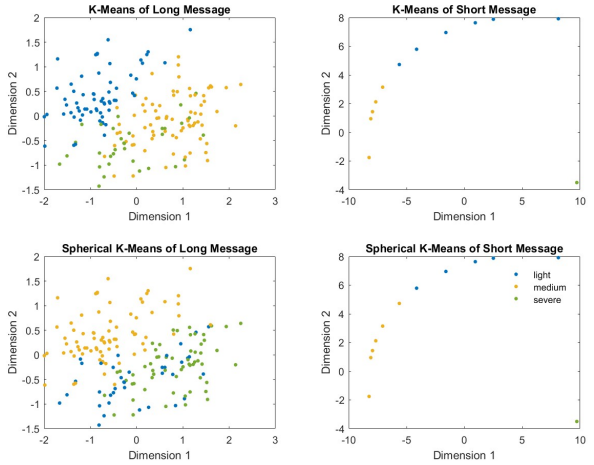

Figure 7. Clustering representation of K-means and Spherical K-means using cosine similarity with PCA.

In particular, the original K-means algorithm showed slightly better performance compared to Spherical K-means in clustering long messages but showed no difference when applied to short messages. These clusters are the most accurate ones achieved based on our experiment with 250 sample posts (chosen for a fair comparison between datasets) for each short messages and long messages based on three

categories of pain-related keywords used by TN patients as follows:

***Light*:** help, year, life, know, many
Medium: pain, face, like, help, would
Severe: disease, suicide, suicide disease

4. 1. Comparison of algorithms with baseline

According to the results of our experiments the generated clusters are comparable to our baseline categories (Table 3) as *Light* cluster contains more words from "Awareness" set, *Medium* cluster contains more words from "Pain experience" set and *Severe* cluster contains more words from "TN Description" set vocabularies.

Table 3. Comparison of clustering results with baseline vocabulary set (Facebook and Twitter)

TN Custer	Sample Post
Light (Awareness)	• "National Organization for Rare Disorders, Inc. (NORD). RARE Diseases Day is coming up soon" • "I have a rare condition called Trigeminal Neuralgia aka The Suicide Disease because of the intense" • "Help for trigeminal neuralgia?"
Medium (Pain experience)	• All I can say is at least you know if you need to go to the hospital because the pain is so bad" • "Oh boy. I've had people say some pretty strange stuff to me too (I have rheumatoid arthritis, …)"
Severe (TN Description)	• "National Multiple Sclerosis Society..Anyone else have Trigeminal Neuralgia! So painful! • "I didn't think much about rare diseases, before I became a patient who has 3 rare neuralgia disorders" • "So much pain if she does have trigeminal neuralgia "

To show how our clustering algorithms performed, we also created three different posts (with similar themes to actual posts from both Facebook and Twitter data) against which we tested our clustering prediction of K-means and Spherical K-means. These posts are based on thematic representation of actual TN posts and are defined for our thematic test as follows:

A) TN Description= *Before it was called a suicide disease. But these days we have several methods that patients can try to alleviate their pain. I think my surgeon could find out what actually was going on and however he said it was not shown clearly on MRI images I took last year. So, it could progress?!!!*

B) Pain experience= *I had this treatment but they gave my dad different meds that helps with pain.*

C) Pain experience= *Same pain experience*

D) Pain experience=*Meds? No I am not taking*

Results of this thematic comparison showed that traditional K-means outperformed Spherical K-means in assigning the posts to their corresponding baseline categories when trained on the Facebook dataset, but both performed poorly when trained on Twitter dataset (Table 4).

Table 4. Comparison of clustering prediction of K-means and Spherical K-means with baseline

Trained on Facebook data
K-means: • (Post A) predicted as→ TN Description • (Post B) predicted as→ Pain experience • (Post C) predicted as→ Awareness • (Post D) predicted as→ Pain experience Spherical K-means: • (Post A) predicted as→ TN Description • (Post B) predicted as→ TN Description • (Post C) predicted as→ TN Description • (Post D) predicted as→ Pain experience
Trained on Twitter data
K-means: • (Post A) predicted as→ TN Description • (Post B) predicted as→ TN Description • (Post C) predicted as→ TN Description • (Post D) predicted as→ TN Description Spherical K-means: • (Post A) predicted as→ Awareness • (Post B) predicted as→ TN Description • (Post C) predicted as→ Awareness • (Post D) predicted as→ Awareness

GSDMM, as a short-text topic modeling algorithm however, showed better performance only when tested on Facebook posts. Below is a summary of GSDMM clustering results for thematic test when compared with Baseline vocabulary set (Table 5).

Table 5. Comparison of GSDMM generated topics and baseline sets

Facebook dataset	Twitter dataset
Topic 1 (Pain experience): pain, neuralgia, would, years, disease, face, many, could, take, time, side, back, medical, never, diagnosed.	*Topic 1 (Awareness)*: trigeminal, treatment, disease, help, teal, via, painful, us, October, face, support.
Topic 2 (Awareness): one, know, suicide, surgery, even, chronic, tn, pain, ms, called, much, well known, us, see, work, please, doctor, thank.	*Topic 2 (Awareness)*: trigeminal, awareness, day, facial, nerve, know, patients, time, hope, migraine, question, helps, still.
Topic 3 *(TN Description)*: trigeminal, neuralgia, get, help, people, nerve, condition, painful, brain, doctors, head, multiple.	*Topic 3 (Awareness)*: pain, suicide, today, neuralgia, please, carbamazepine

For Twitter dataset however, GSDMM did not perform very well as topics generated are not thematically heterogeneous which is expected since Twitter was mostly used for raising awareness about TN as shown by keywords such as awareness, please, today, teal (a selected color for TN) and October 7th (TN awareness day).

5. Discussion and conclusion

In this study, we used unsupervised machine learning algorithms to analyze TN patients' social media posts and derive meaningful structures. Since this exploratory study is the first to consider the analysis of TN social media patient-generated posts, we contend that exploring different methods of clustering beginning with widely known methods will help build fundamental approaches towards advanced cluster and classification analysis of TN and other rare diseases. It can give a sense of what direction should be taken and what other methods must be explored.

According to our experiment results, both K-means and Spherical K-means performed poorly on Twitter and Facebook short message dataset. One reason could be the data does not have enough information belong to clusters within short messages. GSDMM however, could cluster the short dataset (mostly Facebook and those tweets longer than 50 words) with a good trend. One reason is that Facebook data contains more content and is more diverse. Another reason could be that TN patients engage more on Facebook compared to Twitter since they can share more information with no word limits.

This study provides some research implications. First, social media analytics for rare disease is a challenging process in which prior acceptable methods and clustering algorithms may not perform as one expected. We experienced this when applying GDMMS that is well known for handling clustering of very short messages but experienced low performance since our Twitter dataset had mostly less than 50 words and was thematically less heterogenous.

Second, when clustering social media posts about rare disease we may encounter lack of clarity in clusters due to shortage of unified terminology circulated within posts. This could be the case in our dataset in which newcomers and the experienced users used different terms to describe their conditions. More advanced analysis with richer dataset is needed for better clarification. An opportunity that comes out of this exploratory study is to consider utilization of most meaningful clusters and apply classification methods to categorize users into groups based on semantic analysis of posts appeared within each cluster. This will help with the creation of shared vocabulary-based knowledge for TN disorder that is useful, meaningful, easily transferable and communicable among both clinicians and patients. The shared vocabulary creation is common for rare diseases and it is needed for more efficient research, communication and practice.

Third, our approach can also be useful for research aiming at clustering social media data that are not disease specific (i.e. when users posts are not solely based on experience). This can be done for research that is mostly concerned with unlabeled data such as social media marketing, where defining baseline vocabulary sets using techniques such as Bag-Of-Words to derive potential themes (inferred from unique vocabularies) can be a reliable approach to benchmark clusters and topics (e.g. unknown categories of potential customers) that later on may be generated by algorithms.

Finally, our categorization of TN social posts, although high level, when utilized efficiently with advanced topic modeling or classification, can help interventions and research groups to target right participants from the right group of patients.

6. Limitation and future work

Like any other study, this study has some limitations. First, labeling social media data takes time and requires expertise in the problem domain. Hence, we used unlabeled social media data and applied clustering methods whose performance we could evaluate using baseline vocabularies that we inferred

from the datasets. As a result, we could not use evaluation methods suggested for clustering of labeled data (e.g. homogeneity and topic coherence). Future research can address this limitation by using our clusters' *themes* to generate labeled datasets and utilize supervised learning approaches such as classification.

Second, external events such as TN awareness day (October 7th) caused the thematically distributed words, which were about patients' daily concerns (from Pain description cluster), to be clustered as Awareness. Future research can address this limitation by applying topic modelling using algorithms such as Non-negative Matrix Factorization (NMF) to account for potential sub-topics that may derive from these high-level clusters (e.g. Awareness on lack of effective medication).

In the next phase of the research, we plan to experiment with more advance clustering techniques and powerful algorithms such as Density Based Scan (DB-SCAN), Deep Neural Network and Word Embedding using external corpus [51]. The DB-SCAN algorithm can help account for outliers which are common in dealing with social media [52] especially in our case where we experienced high sparsity. In addition, we will consider other dimension reduction techniques and algorithms such as T-SNE for more accurate visual representation of our data. Soft clustering methods such as rough K-means and fuzzy c-means can be explored, and their results can be compared against our clustering results. Moreover, classification algorithms such as bagging and tree-based methods can be used to create accurate and efficient TN categorization models. The ultimate goal of this research is to comply with the Human Phenotype Ontology (HPO) that provides a structured, comprehensive, and well-defined set of terminologies [53] for orofacial rare disease and TN.

7. References

[1] Jones, D.E.J., E. Sturm, and A.W. Lohse, Access to care in rare liver diseases: New challenges and new opportunities. J Hepatol, 2018. 68(3): p. 577-585.

[2] Zakrzewska, J.M. and M.E. Linskey, Trigeminal neuralgia. BMJ, 2014. 348: p. g474.

[3] NHGRI. National Human Genome Research Institute-FAQ About Rare Diseases. 2019; Available from: https://www.genome.gov/27531963/faq-about-rare-diseases/#al-1.

[4] Long, E., et al., An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nature biomedical engineering, 2017. 1(2): p. 0024.

[5] Tait, R.C., M. Ferguson, and C.M. Herndon, Chronic Orofacial Pain: Burning Mouth Syndrome and Other Neuropathic Disorders. J Pain Manag Med, 2017. 3(1).

[6] AANS. Trigeminal Neuralgia. 2018 November, 12]; Available from: https://www.aans.org/Patients/Neurosurgical-Conditions-and-Treatments/Trigeminal-Neuralgia.

[7] Cruccu, G., et al., Trigeminal neuralgia: new classification and diagnostic grading for practice and research. 2016. 87(2): p. 220-228.

[8] Alshukry, A., et al., Trigeminal neuralgia (TN): A descriptive literature analysis on the diagnosis and management modalities. Journal of Stomatology, Oral and Maxillofacial Surgery, 2017. 118(4): p. 251-254.

[9] Allsop, M.J., et al., Diagnosis, medication, and surgical management for patients with trigeminal neuralgia: a qualitative study. Acta Neurochir (Wien), 2015. 157(11): p. 1925-33.

[10] Kitt, C.A., et al., Trigeminal neuralgia: opportunities for research and treatment. Pain, 2000. 85(1): p. 3-7.

[11] Drangsholt, M., E.L. Truelove, and G. Yamuguchi, The case of a 52-year-old woman with chronic tooth pain unresolved by multiple traditional dental procedures: an evidence-based review of the diagnosis of trigeminal neuropathic pain. J Evid Based Dent Pract, 2005. 5(1): p. 1-10.

[12] Melek, L.N., M. Devine, and T. Renton, The psychosocial impact of orofacial pain in trigeminal neuralgia patients: a systematic review. Int J Oral Maxillofac Surg, 2018. 47(7): p. 869-878.

[13] De Toledo, I.P., et al., Prevalence of trigeminal neuralgia: A systematic review. The Journal of the American Dental Association, 2016. 147(7): p. 570-576.e2.

[14] Chicoine, N.H., et al., Surgical Management of Trigeminal Neuralgia in Children. World Neurosurg, 2019. 121: p. 217-221.

[15] Dhar, V.K., et al., Benefit of social media on patient engagement and satisfaction: Results of a 9-month, qualitative pilot study using Facebook. Surgery, 2018. 163(3): p. 565-570.

[16] Wittmeier, K., et al., Analysis of a parent-initiated social media campaign for Hirschsprung's disease. J Med Internet Res, 2014. 16(12): p. e288.

[17] Zakrzewska, J.M., T.P. Jorns, and A. Spatz, Patient led conferences--who attends, are their expectations met and do they vary in three different countries? Eur J Pain, 2009. 13(5): p. 486-91.

[18] Schumacher, K.R., et al., Social media methods for studying rare diseases. Pediatrics, 2014. 133(5): p. e1345-53.

[19] Ventola, C.L., Social media and health care professionals: benefits, risks, and best practices. Pharmacy and Therapeutics, 2014. 39(7): p. 491.

[20] Moorhead, S.A., et al., A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. Journal of medical Internet research, 2013. 15(4): p. e85.

[21] Ranschaert, E.R., et al., Radiologists' usage of social media: results of the RANSOM survey. Journal of digital imaging, 2016. 29(4): p. 443-449.

[22] Kietzmann, J.H., et al., Social media? Get serious! Understanding the functional building blocks of social media. Business horizons, 2011. 54(3): p. 241-251.

[23] Lim, S., C.S. Tucker, and S. Kumara, An unsupervised machine learning model for discovering latent infectious diseases using social media data. J Biomed Inform, 2017. 66: p. 82-94.
[24] Denecke, K. and W. Nejdl, How valuable is medical social media data? Content analysis of the medical web. Information Sciences, 2009. 179(12): p. 1870-1880.
[25] Cheng, Q., et al., Assessing Suicide Risk and Emotional Distress in Chinese Social Media: A Text Mining and Machine Learning Study. J Med Internet Res, 2017. 19(7): p. e243.
[26] Sinnenberg, L., et al., Twitter as a Tool for Health Research: A Systematic Review. Am J Public Health, 2017. 107(1): p. e1-e8.
[27] Neiger, B.L., et al., Evaluating social media's capacity to develop engaged audiences in health promotion settings: use of Twitter metrics as a case study. Health Promot Pract, 2013. 14(2): p. 157-62.
[28] King, D., et al., Twitter and the health reforms in the English National Health Service. Health Policy, 2013. 110(2-3): p. 291-7.
[29] Stone, J., Social media is a lifeline for patients with rare diseases. 2015, Forbes.
[30] Andreu-Perez, J., et al., Big data for health. IEEE journal of biomedical and health informatics, 2015. 19(4): p. 1193-1208.
[31] Zhou, L., et al., Harnessing social media for health information management. Electronic commerce research and applications, 2018. 27: p. 139-151.
[32] Holzinger, A. From Machine Learning to Explainable AI. in 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA). 2018. IEEE.
[33] Svenstrup, D., H.L. Jørgensen, and O. Winther, Rare disease diagnosis: a review of web search, social media and large-scale data-mining approaches. Rare Diseases, 2015. 3(1): p. e1083145.
[34] Tuarob, S., et al., An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. J Biomed Inform, 2014. 49: p. 255-68.
[35] Aggarwal, C.C. and C. Zhai, A survey of text classification algorithms, in Mining text data. 2012, Springer. p. 163-222.
[36] Ramos, J. Using tf-idf to determine word relevance in document queries. in Proceedings of the first instructional conference on machine learning. 2003. Piscataway, NJ.
[37] Surian, D., et al., Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. Journal of medical Internet research, 2016. 18(8): p. e232.
[38] Stier, S., et al., Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter. Political communication, 2018. 35(1): p. 50-74.
[39] Habibabadi, S.K. and P.D. Haghighi. Topic Modelling for Identification of Vaccine Reactions in Twitter. in Proceedings of the Australasian Computer Science Week Multiconference. 2019. ACM.
[40] Higgins, D., et al. An Unsupervised System for Visual Exploration of Twitter Conversations. in Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. 2016.
[41] Hitz, M.-P., et al., Rare copy number variants contribute to congenital left-sided heart disease. PLoS genetics, 2012. 8(9): p. e1002903.
[42] Duwairi, R. and M. Abu-Rahmeh, A novel approach for initializing the spherical K-means clustering algorithm. Simulation Modelling Practice and Theory, 2015. 54: p. 49-63.
[43] Sridhar, V.K.R. Unsupervised text normalization using distributed representations of words and phrases. in Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. 2015.
[44] Sriram, B., et al. Short text classification in twitter to improve information filtering. in Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010. ACM.
[45] Baharudin, B., L.H. Lee, and K. Khan, A Review of Machine Learning Algorithms for Text-Documents Classification. Journal of Advances in Information Technology, 2010. 1(1).
[46] Yin, J. and J. Wang. A dirichlet multinomial mixture model-based approach for short text clustering. in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014. ACM.
[47] Duan, R. and C. Li. An Adaptive Dirichlet Multinomial Mixture Model for Short Text Streaming Clustering. in 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI). 2018. IEEE.
[48] Dunbar, R.I., et al., The structure of online social networks mirrors those in the offline world. Social networks, 2015. 43: p. 39-47.
[49] Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. Proceedings of GSCL, 31-40.
[50] Tenenbaum, J.B., V. De Silva, and J.C. Langford, A global geometric framework for nonlinear dimensionality reduction. science, 2000. 290(5500): p. 2319-2323.
[51] Turian, J., L. Ratinov, and Y. Bengio. Word representations: a simple and general method for semi-supervised learning. in Proceedings of the 48th annual meeting of the association for computational linguistics. 2010. Association for Computational Linguistics.
[52] Ranneries, S.B., et al. Wisdom of the local crowd: Detecting local events using social media data. in Proceedings of the 8th ACM conference on web science. 2016. ACM.
[53] Groza, T., et al., The human phenotype ontology: semantic unification of common and rare disease. The American Journal of Human Genetics, 2015. 97(1): p. 111-124.